Membership Inference Attacks on Deep Regression Models for Neuroimaging

Umang Gupta, Dimitris Stripelis, Pradeep K. Lam, Paul M. Thompson, Jose Luis Ambite, Greg Ver Steeg

Introduction

- Data privacy laws restrict data sharing and aggregation for model training
- In such cases, model sharing is usually allowed, e.g., federated learning
- Can an adversary extract information from only the models and some knowledge of data distribution?

"We show privacy leakage that may be caused by model sharing by demonstrating successful membership inference attacks on neural networks trained for neuroimaging tasks. Membership inference attacks [1] aim to infer if a sample was used to train the model."

Setup

- We attack models trained for Brain age prediction under centralized [2] and federated [3] training setup
- Attacker can access trained model (NN) parameters and has access to some data samples in centralized setup.
- In the federated training setup, we consider attacker to be one of the learners

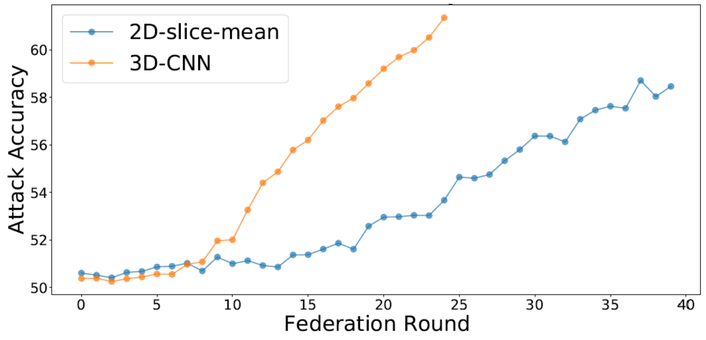
Attack Procedure

- Extract features using the trained model parameters and samples --- gradients, prediction error, activations etc.
- The attacker may train a binary classifier over the data samples.
- Evaluate the accuracy of the attacker on a set that was not seen by the attacker before.

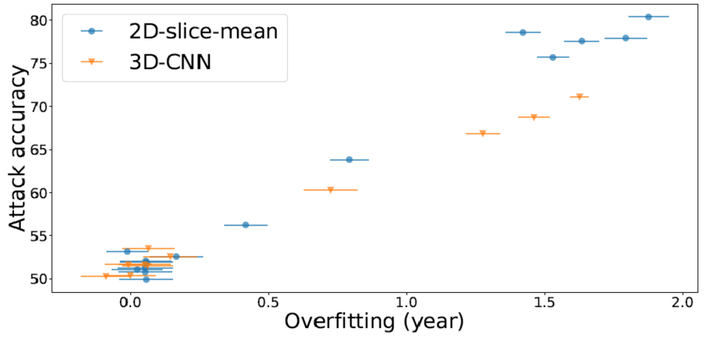
Results

Assumption	2D-Slice-CNN	3D-CNN
Access to training samples	83.04	78.05
Access to training distribution	74.39	71.74

Attack accuracies under different access assumptions for models trained centrally



Attack accuracy increases with federation rounds



Overfitting correlates with Privacy Leakage

References

1] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership Inference Attacks Against Machine Learning Models. In 2017 IEEE Symposium on Security and Privacy (SP), 2017.

[2] Dimitris Stripelis, Jose Luis Ambite, Pradeep Lam, and Paul Thompson. Scaling Neuroscience Research using Federated Learning. In IEEE International Symposium on Biomedical Imaging (ISBI), 2021

[3] Umang Gupta, Pradeep Lam, Greg Ver Steeg, and Paul Thompson. Improved Brain Age Estimation with Slice-based Set Networks. In IEEE International Symposium on Biomedical Imaging (ISBI), 2021.

Information Sciences Institute

Link: https://openreview.net/forum?id=8lL_y9n-CV
Contact: umanggup@usc.edu

Keck School of Medicine of USC

USC Viterbi School of Engineering